IMPERFECT IMAGANATION: IMPLICATIONS OF GANS EXACERBATING BIASES ON FACIAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we show that popular Generative Adversarial Networks (GANs) exacerbate biases along the axes of gender and skin tone when given a skewed distribution of face-shots. While practitioners celebrate synthetic data generation using GANs as an economical way to augment data for training data-hungry machine learning models, it is unclear whether they recognize the perils of such techniques when applied to real world datasets biased along latent dimensions. Specifically, we show that (1) traditional GANs further skew the distribution of a dataset consisting of engineering faculty headshots, generating minority modes less often and of worse quality and (2) image-to-image translation (conditional) GANs also exacerbate biases by lightening skin color of non-white faces and transforming female facial features to be masculine when generating faces of engineering professors. Thus, our study is meant to serve as a cautionary tale.

1 INTRODUCTION

The use of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) has grown significantly and due to data-demand of deep learning models, when faced with sparse data (owing to paywalls, privacy concerns, etc.) practitioners often turn to promising data augmentation solutions. While earlier computer vision works focused on performing affine transformations to existing samples (O'Gorman & Kasturi, 1995; Bloice et al., 2017), using GANs for synthetic data generation has recently become popular (Teich, 2019; Nisselson, 2018). GANs generate such data by approximating the original distribution with a limited training set and create examples that appear novel. These examples give a (false) sense of sampling unseen data from the same underlying distribution as the original training data, making GANs a seemingly perfect candidate for data augmentation. We note that even this best-case scenario would be a territory for practitioners to tread lightly; GAN-generated data for augmentation would only propagate the existing biases of the real-world data. Owing to theoretical limitations of GANs (Arora & Zhang, 2017), we show a grim reality: the generated data learns a distribution shifted from that of the real world, one which exacerbates these biases and disproportionately underrepresents those already in the minority, both in number and quality. This poses serious ethical implications on any downstream tasks trained on a synthetically-augmented dataset, especially when biases exist along protected or embargoed attributes.

2 ARCHITECTURE AND APPROACH

Mode Collapse GANs are known to estimate an equilibrium of a minimax game played by a generator network G and discriminator network D. While D, a binary classifier, learns to discriminate between images that come from a real-world data distribution p_{data} and those that do not, G learns to generate images from p_{GAN} and fool D into classifying them as coming from p_{data}. In the presence of infinite training data, computation time and network capacity for the generator and the discriminator, this process ensures that the p_{GAN} distribution generated by G converges to that of the training data p_{data} (Goodfellow et al., 2014). In reality, GAN-generated distributions are not nearly as diverse as their training distributions (Arora & Zhang, 2017; Arora et al., 2017) and the support (i.e. possible feature combinations of the generated data) is only representative of a small subset of what one would expect to see when sampling data from the real distribution. The support size of the generated images is constrained by the capacity of D. G collapses because the set of

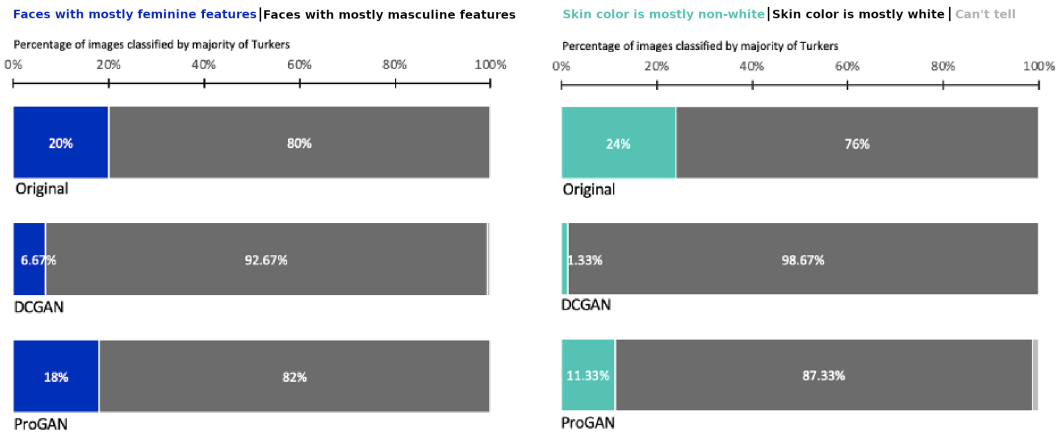

Figure 1: Distribution of human classifications on gender and skin color.

noise inputs that would correspond to some minority mode in the image space has (by definition) a low probability of being seen by D. As G only optimizes its own weights over the feedback from D, it rarely learns to generate these modes (Che et al., 2016).

There are several related works– Zhao et al. (2018) studies GANs' bias and generalization to unseen modes without discussing the problem of GANs collapsing to existing modes. While mode collapse is a well-studied phenomenon (Grnarova et al., 2018; Goodfellow, 2016; Che et al., 2016; Arora et al., 2017) and several GAN variants have been developed to alleviate its effects (Metz et al., 2016; Srivastava et al., 2017; Arjovsky et al., 2017; Miyato et al., 2018; Tolstikhin et al., 2017; Karras et al., 2017), a distinction is rarely made between uniform and non-uniform training datasets. On these lines, Mishra et al. (2018) empirically shows that the divergence between p_{data} and p_{GAN} does indeed worsen as the training data is more skewed, however using four scalar metrics which do not offer much insight on *how* the distributions differ. For a dataset that is biased along latent axes (e.g. gender and skin color), we hypothesize G (we try several GAN variants) collapses to modes in the majority groups (e.g. masculine and white faces) amplifying biases that exist in the original data.

Data Collection and Processing To test our hypothesis, we construct a dataset of faces of engineering professors from U.S. universities that are (1) listed in the top 47 of US News' most recent "Best Engineering Schools" and (2) had public access to faculty directories with images. The data exhibits bias along the latent dimensions of gender and race and thus, is an appropriate test-bed to study the amplification of bias in GAN-based data generation. We gather a total of $17,245$ engineering faculty 64×64-pixel headshots using an unsupervised face detector (Dalal & Triggs, 2005).

3 Experiment and Results

To explore the diversity of p_{GAN} we test the performance on three GANs (1) DCGAN (Radford et al., 2015): the most common GAN used by practitioners due to its minimal requirements for compute power and off-the-shelf availability (carpedm20, 2015), (2) ProGAN (Karras et al., 2017): a state-of-the-art GAN for sample quality and known to addresses the mode-collapse problem and to overcome the quality-variance tradeoff (Karras et al., 2019; 2020), and (3) CycleGAN (Zhu et al., 2017): the most well-known image-to-image translation GAN, which transforms an image from one domain to another by minimizing cycle-consistency and identity losses. We show experiments on two other GAN architectures designed to address mode collapse – Wasserstein GAN (Arjovsky et al., 2017) and AdaGAN (Tolstikhin et al., 2017) – in the appendix.

3.1 Imagining Engineers from Scratch

We assess the data from the GAN variants – DCGAN and ProGAN – by asking humans to annotate images from the original and generated datasets along the dimensions of race and gender. To account for variance in model training, we generate 50 images from three seeds where each seed trains

Figure 2: Illustrative test set of transformations on non-white (two rows on the left) and female celebrities (two rows on the right). Original and stylized images are one atop another respectively.

the DCGAN and the ProGAN for 50 epochs. We conduct 4 seven-minute human study tasks in a between-subject design fashion (each annotator saw images belonging to only one set) and leveraged data from 234 master Turkers on Amazon's MTurk platform. Each worker performed the tasks:

[`T1 (a/b)`] Human subjects were asked to select the most appropriate option for an image x sampled from [`T1a`] p_{data} and [`T1b`] $G(z)$ with the following options: (1) face mostly has masculine features, (2) face mostly has feminine features, and (3) neither of the above is true.

[`T2 (a/b)`] Human subjects were asked to select the most appropriate option for an image x sampled from [`T2a`] p_{data} and [`T2b`] $G(z)$ from the list of following options: (1) skin color is non-white, (2) skin color is white, and (3) can't tell.

We presented each annotator with 52 images– 50 from the original/generated data and two high quality trivial images with known labels for gender and skin color. This helped us prune 18 bad datapoints. We had 30 valid data points for all generated datasets and 25 for the original distribution. We considered majority-voting to categorize an image as belonging to a class. Figure 5 in the appendix contains the resulting charts.

3.1.1 Results

We plot the results for `T1a` and `T1b` in Figure 1 (left) and find that (1) both DCGAN and ProGAN penalize the original 20% of images with mostly feminine features being DCGAN the most penalizing, reducing the percentage to 6.67%. A one-tailed two-proportion z-test yields a p-value of 0.0032 confirming the amplification of bias across the latent dimension of gender for DCGAN and (2) for tasks `T2a` and `T2b` (Figure 1, right) the proportion of non-white faces decreased from 24% in the original dataset to 1.33% in the DCGAN-generated dataset and to 11.33% for ProGAN. The p-value obtained (2.7×10^{-8} for DCGAN and 1.05×10^{-3} for ProGAN) show strong statistical significance as both GANs collapse along the latent dimension of race, biasing the synthetic faces toward lighter skin tones. Note that while ProGAN did not collapse along the axis of gender, it was not immune to collapsing along the axis of other protected features (eg. skin color). We notice that the synthetic data not only propagates but exacerbates those biases against minority populations.

Quality and Confidence Metrics We measure the consensus among Turkers by the amount of votes needed to classify each image in the axes of gender and color. For DCGAN, we find that the proportion of images labelled as non-white and female decreases as the voting threshold increases. This is indicative of a higher level of agreement between participants and shows that the quality of generated images for the minority classes is worse than that of the majority classes. ProGAN does not exhibit this disparity in quality across gender, but it produces lower quality for non-white faces than white ones.

3.2 Imagining Engineering Counterparts

As image-to-image translation GANs' output distributions conditioned on the input, our intuition was that they may be less susceptible to exacerbating biases. For instance, in our task where gender is a latent feature and feminine faces are underrepresented, a GAN, provided with the input image of a female, would have to actively convert it into a male one. Unfortunately, it is known that even these conditional GAN variants are not immune to mode collapse (Ma et al., 2018). However, how conditional variants of GANs react to sensitive social features such as race and gender remains an open question.

To study this, we train a CycleGAN (Zhu et al., 2017) to stylize faces of non-engineering professors to look like engineering faculty. Thus, our target/output domain consists of the engineering faculty face dataset leveraged in the previous experiment and our input domain is the CelebA dataset (Liu et al., 2015) consisting of over 200,000 annotated images of celebrities. As our dataset consists of only 16,500 images, we randomly sample 16,500 faces from the CelebA dataset for training. We then create a held-out test set from CelebA in which we have 100 images for each of the four categories– white, non-white, male, and female.

In Figure 2, we showcase the transformation of celebrity faces that are representative of the minority categories (i.e. non-white, female) in the engineering professors dataset. While we see that the GAN learns to add glasses or creating smiling expressions, not all the modifications learned are socially harmless, we also notice that it lightens the skin tone of non-white celebrities and imparts masculine aspects to the faces of female celebrities. While it is reasonable to expect that a GAN might perpetuate and exacerbate biases along any arbitrary dimension where there exists a skew in the training set, we stress that this kind of innocuous bias is not our focus. Machine learning systems are designed to find correlations to recognize patterns, but this correlation-seeking becomes problematic for social features when models perpetuate and exacerbate biases for minority groups who have faced systemic disadvantage or discrimination. Before concluding, we highlight a case-study where such models are having adverse real-world impact.

4 Real-World Applications and Conclusion

While our experiments meant to serve as example, the bias-exacerbation consequences of mode collapse in GANs can be seen in real-world applications. Snapchat, a popular image-sharing platform, has recently taken advantage of the image-to-image translation capabilities of conditional GANs such as CycleGAN for their "My Twin" lens, according to several sources (Yanjia Li, 2020; Magazine, 2019; Jang, 1970; red, 2019). We show that this presumably conditional-GAN-based technology reacts to the sensitive features this work discusses. When applying this lens to a female face, the GAN should ideally make no changes, but when used on women of color, it lightens skin tone, though this is not the case for white women using the same filter. While we have not performed a comprehensive study, the observations and claims open an intriguing research problem (Baeza-Yates, 2016). Examples of the lightened complexions on women of color and white women can be seen in Figure 6 in the appendix.

The implications of using a biased facial dataset augmented via GANs for a downstream task could be severe. The use of machine learning models on facial data is already prevalent in critical decision-making scenarios such as employment (Hymas, 2019), healthcare (Bahrampour, 2014), education (Kaur & Marco, 2019), criminal justice (Harwell, 2019), as well as security innovations like deep-fake detection (Dolhansky et al., 2019). It is of clear ethical import that we ensure our training sets and models are fair and diverse with respect to sensitive features. At the very least, they ought not to rig the system *against* already underrepresented minorities.

GANs have proven to create less diverse distributions than the original they are trained on, but the implications of mode collapse remain unclear in scenarios where the training distribution p_{data} is biased toward certain feature values (eg. males) along a latent feature (eg. gender). To study this, we empirically show how GANs trained on a demographic already skewed toward white and male faces exacerbate social biases in the generated distribution p_{GAN}. In our setting, mode collapse occurs on a majority latent mode of the original data and causes a severe under-representation of feminine facial features and non-white skin tones in the generated dataset. We also demonstrate that this perpetuation of biases against female and non-white features occurs in image-to-image translation GANs, first stylizing celebrities' faces to look like those of engineering professors, and next by conducting a case study on Snapchat's "My Twin" lens. Beyond implications about social issues, this work should serve as a general caution against using GAN-based data augmentation techniques to alleviate problems arising from sparse or unbalanced datasets for any downstream task. There seems to exist a false sense of security that GANs can generate *novel* data samples which pick the expected semantic features relating to the defect, and place them in previously unseen settings. In actuality, the augmented data might be underrepresenting or compromising image quality for some crucial feature of the real-world data.

REFERENCES

r/machinelearning - [d] is the new snapchat gender filter gan-based? https://www.reddit.com/r/MachineLearning/comments/bo4orw/d_is_the_new_snapchat_gender_filter_ganbased/, 2019.

Julia Angwin and Jeff Larson. Machine bias. https://bit.ly/37uNrV2, 2016.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Sanjeev Arora and Yi Zhang. Do gans actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 224–232, 2017.

Ricardo Baeza-Yates. Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science*, pp. 1–1, 2016.

Tara Bahrampour. Can your face reveal how long you'll live? new technology may provide the answer. https://wapo.st/2GrQKjL, Jul 2014.

Marcus D Bloice, Christof Stocker, and Andreas Holzinger. Augmentor: an image augmentation library for machine learning. *arXiv preprint arXiv:1708.04680*, 2017.

carpedm20. DCGAN implementation in Tensorflow. https://github.com/carpedm20/DCGAN-tensorflow, 2015.

Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pp. 886–893, 2005.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

Timnit Gebru. Oxford handbook on ai ethics book chapter on race and gender. *arXiv preprint arXiv:1908.06165*, 2019.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Paulina Grnarova, Kfir Y Levy, Aurelien Lucchi, Nathanael Perraudin, Thomas Hofmann, and Andreas Krause. Evaluating gans via duality. *arXiv preprint arXiv:1811.05512*, 2018.

Drew Harwell. Oregon became a testing ground for amazon's facial-recognition policing. but what if rekognition gets it wrong? https://wapo.st/2TXiXHl, Apr 2019.

Yuta Hiasa, Yoshito Otake, Masaki Takao, Takumi Matsuoka, Kazuma Takashima, Aaron Carass, Jerry L. Prince, Nobuhiko Sugano, and Yoshinobu Sato. Cross-modality image synthesis from unpaired data using cyclegan: Effects of gradient consistency loss and training data size. pp. 31–41. Springer Verlag, 2018.

Sheng-Wei Huang, Che-Tsung Lin, Shu-Ping Chen, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai. Auggan: Cross domain adaptation with gan-based data augmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 718–731, 2018.

Charles Hymas. Ai used for first time in job interviews in uk to find best applicants. https://bit.ly/2vmBwu9, Sep 2019.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Eric Jang. Fun with snapchat's gender swapping filter. https://blog.evjang.com/2019/05/fun-with-snapchats-gender-swapping.html, Jan 1970.

junyanz. CycleGAN. https://github.com/junyanz/CycleGAN, 2017.

junyanz. CycleGAN and pix2pix in PyTorch. https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix, 2018.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Harmeet Kaur and Tony Marco. A new york school district is bringing in facial recognition software. rights groups say it could spell trouble for students. https://cnn.it/30WLaiX, May 2019.

Sayeri Lala, Maha Shady, Anastasiya Belyaeva, and Molei Liu. Evaluation of mode collapse in generative adversarial networks. *High Performance Extreme Computing*, 2018.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5657–5666, 2018.

Paper Magazine. The dark implications of facial swap filter technology. https://bit.ly/2vrmjbv, Sep 2019.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2437–2445, 2020.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Deepak Mishra, Aravind Jayendran, Varun Srivastava, Santanu Chaudhury, et al. Mode matching in gans through latent space learning and inversion. *arXiv preprint arXiv:1811.03692*, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Evan Nisselson. Deep learning with synthetic data will democratize the tech industry. https://tcrn.ch/2RRUAYP, 2018.

Lawrence O'Gorman and Rangachar Kasturi. *Document image analysis*, volume 39. IEEE Computer Society Press Los Alamitos, 1995.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):1–9, 2019.

Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pp. 3308–3318, 2017.

David A. Teich. Synthetic data is a tool for improving training and accuracy of deep learning systems. https://bit.ly/36moolG, 2019.

Ilya O Tolstikhin, Sylvain Gelly, Olivier Bousquet, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. Adagan: Boosting generative models. In *Advances in Neural Information Processing Systems*, pp. 5424–5433, 2017.

Ethan Yanjia Li. Gender swap and cyclegan in tensorflow 2.0 – ethan yanjia li, 2020. URL `https://yanjia.li/gender-swap-and-cyclegan-in-tensorflow-2-0/`.

Shengjia Zhao, Hongyu Ren, Arianna Yuan, Jiaming Song, Noah Goodman, and Stefano Ermon. Bias and generalization in deep generative models: An empirical study. In *Advances in Neural Information Processing Systems*, pp. 10792–10801, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Figure 3: Images of professors generated by popular GAN architectures trained on our engineering professors dataset; WGAN, AdaGAN, and ProGAN attempt to address the mode-collapse problem.

A Appendix

In addition to the variants mentioned in the paper (DCGAN and ProGAN), we investigate the performance of another two GANs which claim to reduce mode collapse: *Wasserstein* GANs or WGAN (Arjovsky et al., 2017) and AdaGAN Tolstikhin et al. (2017).

A.1 Studies with Commercially Available Gender Classification Systems

As a less subjective approach to labelling, we use Microsoft Azure's Face API for classifying 5000 images from the training set and 5000 generated by the three different variants of GAN: the popular DCGAN and two others that attempt to address the problem of mode collapse, AdaGAN and ProGAN (we omit WGAN due to poor image quality). To ensure our results are not specific to a single generation, we obtain 5000 images by sampling from three runs of each GAN with different random seeds for weight initialization. We show 100 images randomly sampled from these 5000 images obtained from each GAN variant in Figure 3.

In Figure 4, we show the percentage of images classified as female, male and "can't tell." We perform a one-tailed two-proportion z-test on the original and generated sample for the proportion of females to assess the null hypothesis that the proportion of feminine features in the synthetic distribution for all GAN variants is equivalent to the proportion of those in the original distribution.

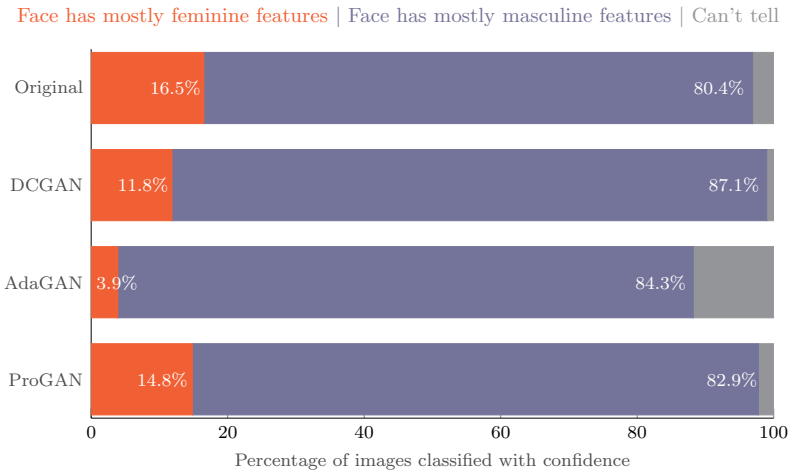

Figure 4: The percentage of faces classified as female, male and can't tell by Microsoft Azure's Face API decreases from 16.5% in the original dataset significantly in the synthetically generated datasets across several GAN variants that are popular or attempt to address the mode collapse problem.

In the original dataset, 16.5% are labelled as females while 80.4% are labelled as male, clearly indicating a bias this dataset of engineering professors. DCGAN exacerbates this bias significantly (with a p-value of 3.2×10^{-6}), bringing down the percentage of females in the generated set to 11.8%. The quality of images generated by AdaGAN is significantly worse than the ones produced by all other variants as indicated by the spike in the number of images, from 3.1% in the original dataset to 11.8% (with a p-value of 3.0×10^{-53}). Surprisingly, regardless of the poor quality, there is a significant increase in the number of generated images that are classified as male (from 80.4% in the original data to 84.3%) while the number of generated images that are classified as females has a substantial drop (from 16.5% to 3.9%). This also highlights that many of the other GAN variants that seek to address mode-collapse but have shown to be worse than AdaGAN (Lala et al., 2018) such as WGAN (Arjovsky et al., 2017), VEEGAN (Srivastava et al., 2017), Unrolled GAN (Metz et al., 2016) either affect the quality of generated images, exacerbate the biases over latent features such as gender, or both. On the other hand, the more recent architecture ProGAN, clearly outperforms both the popular DCGAN and the AdaGAN in terms of reducing the exacerbation of bias and image quality. It only decreases the percentage of females in its generated set by 1.7%, even though this is a significant exacerbation of bias along the latent dimension of gender (with a p-value of 0.09008). Our results show that popular and state-of-the-art GAN variants paints an optimistic picture of this technology for data-augmentation while suffering from problems of exacerbating biases along latent dimensions.

A.2 Confidence Metrics for DCGAN and ProGAN

We measure the confidence of the annotators for the synthetic datasets by plotting how they would be classified if we were to use a different thresholding technique than majority voting. In Figure 5, we show what the classifications of the images are with different voting thresholds. The x-axis represents the number of votes required to classify an image, and the y-axis represents the proportion of images from the dataset which would have been classified as having a certain label with that voting threshold. This is a metric for consensus among the group of annotators that the images belong to a certain class. For the original data, roughly the same proportion of images are classified as male, female, white, and nonwhite irrespective of the number of Turkers needed to vote. In other words, the Turkers are confident about which faces are male, female, white, and non-white. This is not the case for the synthetic distributions. For DCGAN, the proportion of images that are marked as female or non-white significantly drops as it requires more Turkers to vote for that label, and they only lose confidence over the images depicting the minority gender and race; the proportion of images marked as male or white does not drop as the voting threshold increases. For ProGAN's images, the Turkers are confident about male, female, and white faces, but not about non-white faces.

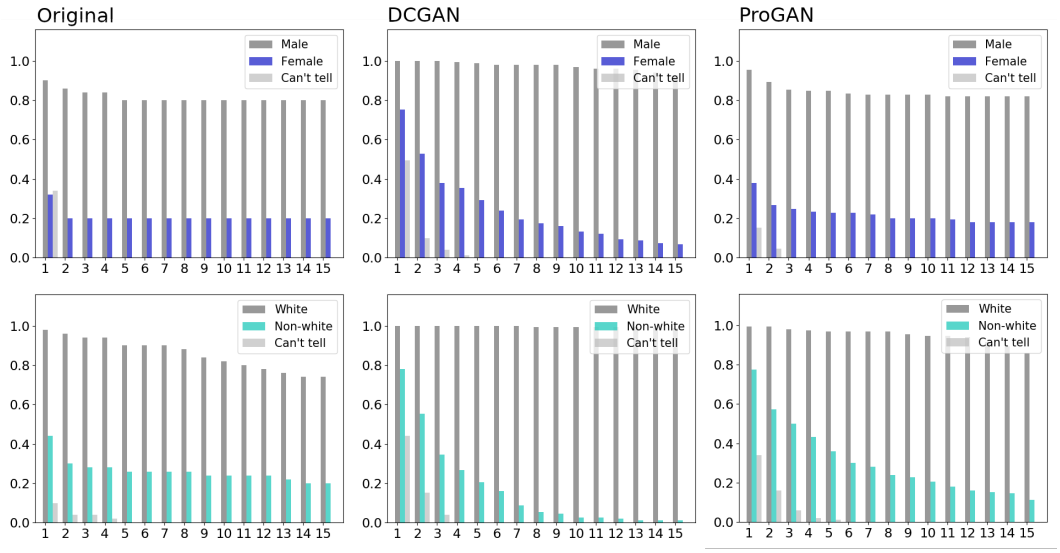

Figure 5: Human annotator agreements on skin color and gender between professor headshots from the original and synthetic (generated by DCGAN and ProGAN) distributions. The number of images labeled as masculine, feminine or neither, changes as the threshold number of votes required to categorize an image into a particular category increases from 1 to 15.

A.3 SNAPCHAT CASE STUDY

Image-to-image translation GANs, such as pix2pix or CycleGAN (Isola et al., 2017; Zhu et al., 2017) adjust colors and textures in an already-existing image from some input domain to map it to another class. Normally, the input and target domains are closely related and the mapping can be achieved by changing the geometries minimally, if at all. Some examples of successful applications for image-to-image translation are conversion of horses to zebras, street photographs to their semantic segmentation, aerial photos to Google maps, and summer landscapes to winter landscapes. CycleGAN is the most popular off-the-shelf GAN variant used by machine learning practitioners today, as measured by the number of stars on the most-used GitHub repositories for this model (junyanz, 2018; 2017), and has also, predictably, been a popular choice for synthetic data augmentation (Hiasa et al., 2018; Sandfort et al., 2019; Huang et al., 2018). Just as with the unconditional variants, our motivation is to explore if and how the diversity of the generated distribution p_{GAN} for the target distribution differs from the training distribution p_{data}.

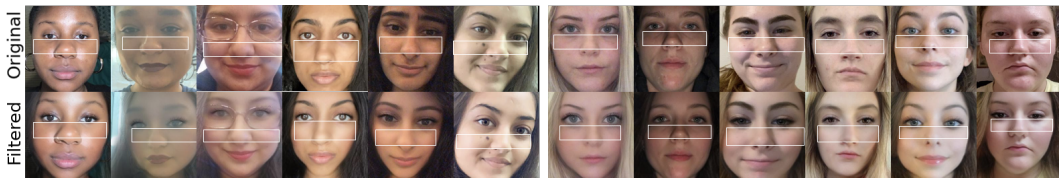

Figure 6: Faces of women of color (left six columns) and white women (right six columns) before and after using Snapchat's female gender face lens, top and bottom respectively. The sections used for the skin-color machine analysis are highlighted in white.

To assess how the skin color changed between pairs of images objectively, we crop a section of the face under the eyes and above the tip of the nose, spanning both cheeks, and find its average pixel value, then we map the RGB vector, using L2-norm distance, to the closest standard shade in the L'Oréal skin color chart[1]. While not considering skin warmth, only skin lightness, we show that the lens lightens non-white faces by one shade consistently for five faces and produces no effect for one

[1] https://www.loreal.com/en/articles/science-and-technology/expert-inskin/

of them. On the other hand, it performed randomly for white faces in our example, lightening two by one shade, darkening two by one shade, and not affecting two. A potential cause of lightening skin tones in women of color is that a GAN used by the face lens collapses all inputs in a region of the image space to output lighter colors. However, more rigorous studies should be performed to make certain claims. Our case study offers initial support for the narrative of Snapchat's beautification face lenses lightening skin tones for people of color.

A.4 Downstream Tasks and Vulnerable Communities

The glaring ethical problem with automated, machine-learning powered tools is that they are "most often used on people towards whom they exhibit the most bias," and that the errors arising from bias "can be much more costly for those in marginalized communities than other groups" (Gebru, 2019). Classification tasks in the real world suffer from this dilemma. In criminal justice, automated tools predict recidivism risk in a system which disproportionately punishes Black and Hispanic people. It is unfortunate yet unsurprising, then, that the risk assessment software used in state criminal justice systems is biased against Black people (Angwin & Larson, 2016). The classification system is given input with over 137 features – not including race – and disproportionately classifies Black defendants as medium or high risk. In employment, automated tools predict candidate performance and fit in industries which are already male-dominated. A hiring system designed by Amazon in 2018 faced public backlash when it was found to discriminate against female candidates by penalizing résumés which included participation in women's organizations. The classification system scraped résumés of candidates from the past ten years and was never given gender as an input feature. Classifiers are not the only automated tool who would use data generated by GANs. In 2020, PULSE (Menon et al., 2020), a face "depixelizer," received widespread backlash on social media (especially from world-famous contributors to the field of AI ethics) because it was shown to upsample images of non-white faces to have Caucasian features. The authors of this paper perform several studies in response and conclude that the biases in PULSE derive directly from the biased performance of the GAN from which it receives generated data. The major takeaway of all discussions mentioned here is that the data bias problem cannot be reduced solely to the dataset used. It seems that popular automated data generation tools, namely GANs, will not merely perpetuate the patterns found in the data (the theoretical ideal for the technologies), but rather amplify them. The question is, then, how we can regulate the societal applications for which these known flawed systems are used.